Google Data Analytics Capstone Project (Cyclistic)

Analysis of Cyclistic, a successful bike share company based in Chicago.

This project follows the successful conclusion of my Google Data Analytics Professional Certificate. It is an excellent opportunity to show the skills acquired throughout the course, such as cleaning, analysing and presenting data.

As a work methodology, I followed the data analysis process steps: **ask, prepare, process, analyse, share, and act.**

SCENARIO

Cyclistic is a bike-share program that was launched in 2016 and has grown into a successful venture with a fleet of **5,824 bicycles** across **692 stations in Chicago**. The program offers a range of pricing plans, including single-ride passes, full-day passes, and annual memberships. The finance team has identified that annual members generate higher profits than casual riders, prompting the company to focus on converting casual riders to annual members as a key strategy for future growth.

To this end, Cyclistic's marketing team aims to analyse historical bike trip data to identify trends that can inform the development of effective marketing strategies. The team seeks to understand the differences between annual members and casual riders, as well as the factors that might motivate casual riders to purchase an annual membership. Additionally, they will explore how digital media can be leveraged to optimise marketing tactics for maximum impact.

Overall, the objective of this initiative is to design marketing strategies that will increase the number of annual members, thus driving future growth for Cyclistic.

ROLE

As a Junior Data Analyst within Cyclistic’s Marketing Analyst Team, I am responsible for analysing historical bike trip data in order to understand the differences between casual riders and annual members.

1. ASK

Business Question: How do annual members and casual riders use Cyclistic bikes differently?

Business Task: Design a marketing strategy that converts casual riders into annual members

Key Stakeholders: My key stakeholders are the Marketing Director, **Lily Moreno,** and the Cyclistic Executive team.

2. PREPARE

Data Source: The data has been made available and licenced by Motivate International Inc and can be found here. It contains historical data from March 2022 to March 2023.
The data was downloaded and stored on my hard drive. I did a backup of the original data.

Data Organization: Each Excel file contains 13 columns with data such as ride id, rideable type, location and so on.

Limitations: "Data-privacy issues prohibit you from using riders' personally identifiable information. This means that you won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes"

3. PROCESS

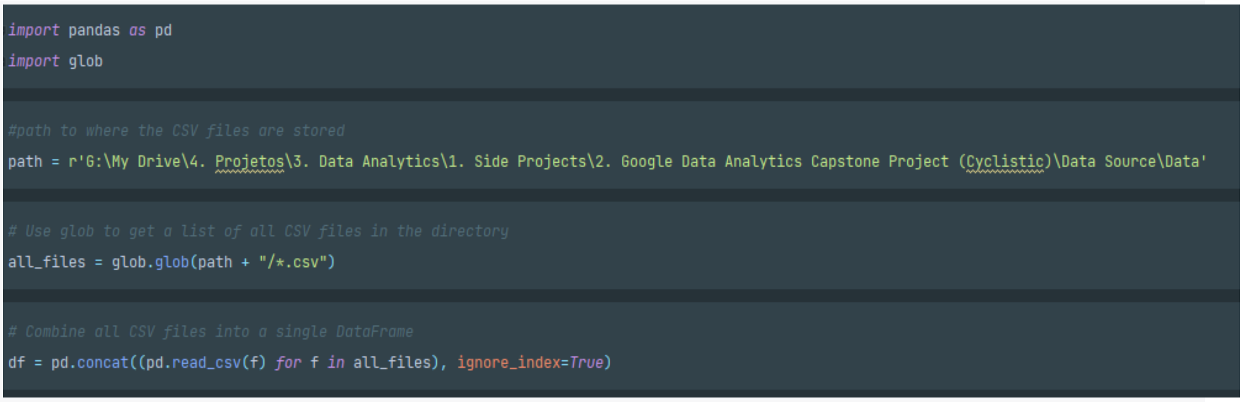
First Approach

I used Microsoft Excel to get a first look at the available data. After checking each file I was able to conclude that the data was consistent across all files, which means, all have the same number of columns and similar nomenclature.

Combine the data

Due to the large number of rows, I choose to work with **Python** in **PyCharm.**

In this analysis, I imported the necessary libraries (pandas and glob) and set the path to the folder containing the CSV files. Then, I used the glob module to get a list of all CSV files in the directory and combined them into a single DataFrame using the pandas concat() function.

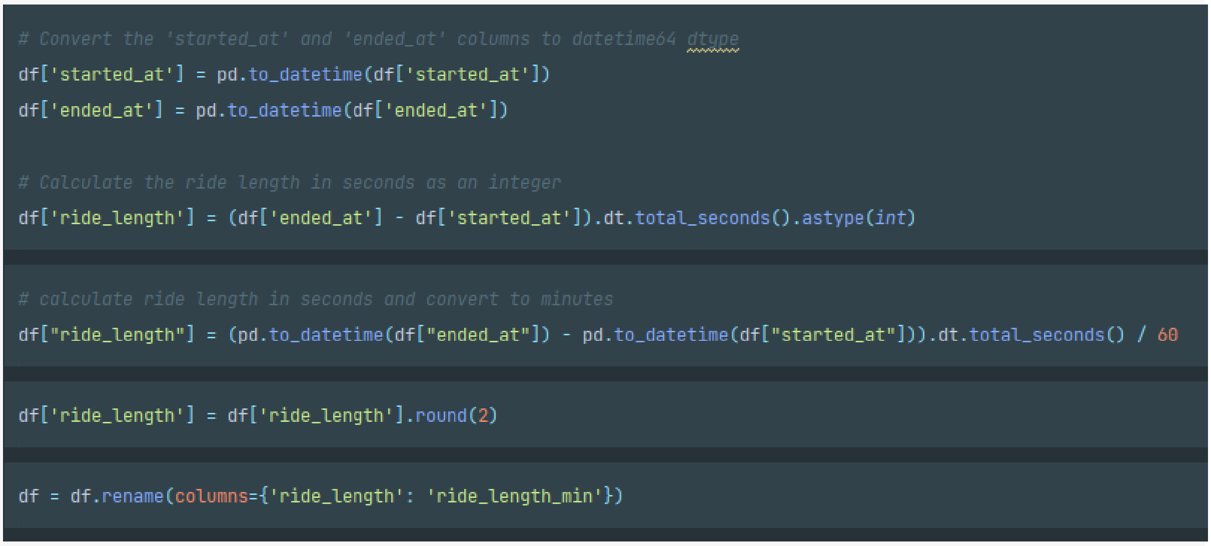


```
import pandas as pd
import glob

#path to where the CSV files are stored
path = r'G:\My Drive\4. Projetos\3. Data Analytics\1. Side Projects\2. Google Data Analytics Capstone Project (Cyclistic)\Data Source\Data'

# Use glob to get a list of all CSV files in the directory
all_files = glob.glob(path + "/*.csv")

# Combine all CSV files into a single DataFrame
df = pd.concat((pd.read_csv(f) for f in all_files), ignore_index=True)
```

New column

Creating a column called "ride_length" allows me to have a quantitative measure of the duration of each ride. This column can be useful in analysing the average ride length, identifying patterns in ride lengths based on factors such as day of the week or time of day, and identifying outliers or extreme values in the data. It can also be helpful in visualising the data and identifying trends over time.

To do this, I converted the “started_at” and “ended_at” columns to datetime64 data type and calculated the ride length in seconds as an integer. Then, I calculated the ride length in minutes and rounded it to two decimal places. I also renamed the column to 'ride_length_min'

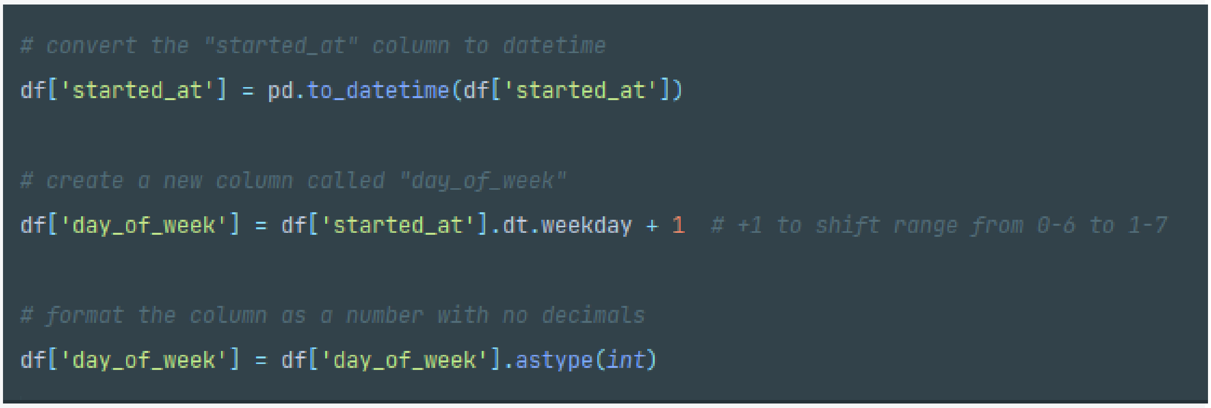


```
# Convert the 'started_at' and 'ended_at' columns to datetime64 dtype
df['started_at'] = pd.to_datetime(df['started_at'])
df['ended_at'] = pd.to_datetime(df['ended_at'])

# Calculate the ride length in seconds as an integer
df['ride_length'] = (df['ended_at'] - df['started_at']).dt.total_seconds().astype(int)

# calculate ride length in seconds and convert to minutes
df["ride_length"] = (pd.to_datetime(df["ended_at"]) - pd.to_datetime(df["started_at"])).dt.total_seconds() / 60

df['ride_length'] = df['ride_length'].round(2)

df = df.rename(columns={'ride_length': 'ride_length_min'})
```

New column

Creating the column "day_of_week" allows me to analyse the day of the week that each ride occurs. This information can be useful for identifying patterns in customer behaviour, such as whether there are certain days of the week with higher demand for bike rentals. Additionally, I can use this information to optimise pricing and marketing strategies based on the day of the week.

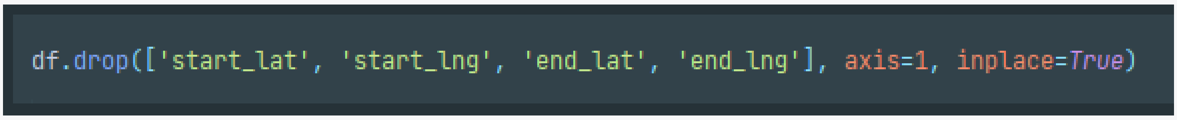
To do this, I created a new column called 'day_of_week' and calculated the day of the week that each ride started using the weekday() method in pandas. I shifted the range from 0-6 to 1-7 to match the format where 1 represents Monday and 7 represents Sunday. Finally, I formatted the 'day_of_week' column as a number with no decimals.



```
# convert the "started_at" column to datetime
df['started_at'] = pd.to_datetime(df['started_at'])

# create a new column called "day_of_week"
df['day_of_week'] = df['started_at'].dt.weekday + 1  # +1 to shift range from 0-6 to 1-7

# format the column as a number with no decimals
df['day_of_week'] = df['day_of_week'].astype(int)
```

Drop unnecessary columns

I dropped the 'start_lat', 'start_lng', 'end_lat', and 'end_lng' columns since they were not relevant to the analysis.

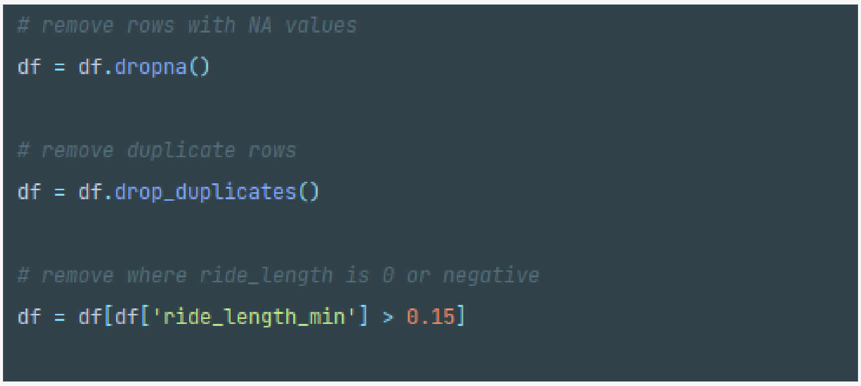


```
df.drop(['start_lat', 'start_lng', 'end_lat', 'end_lng'], axis=1, inplace=True)
```

4. Analyze

Data Cleaning

To ensure that the data is clean and accurate, several data-cleaning steps were taken. Firstly, any rows with missing values were removed from the dataset using the "dropna" function. Then, any duplicate rows were also removed to avoid any possible biases in the analysis. Finally, any rows where the ride length was either 15 seconds or lower e were removed from the dataset to ensure that only valid ride data was included in the analysis.

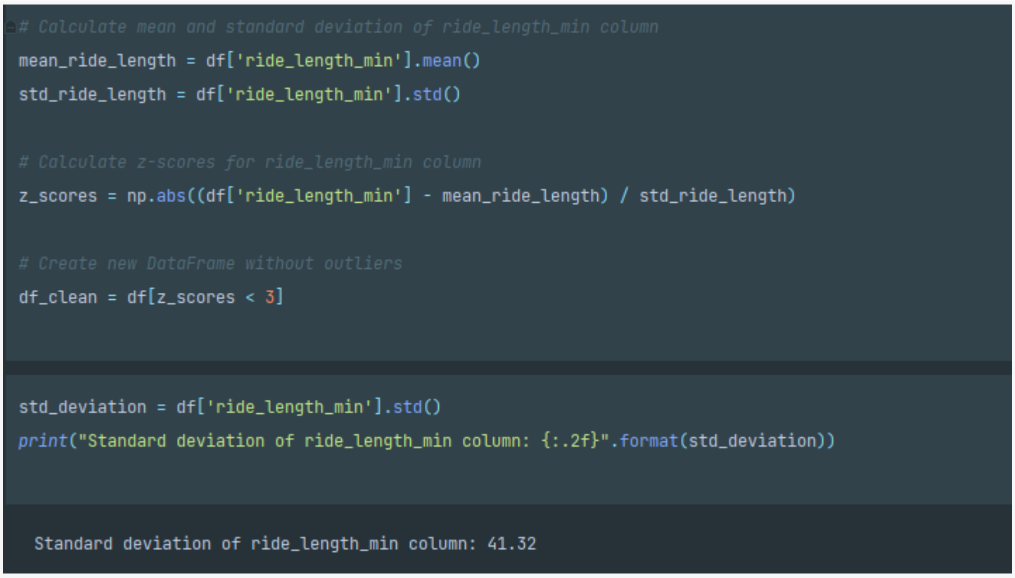


```
# remove rows with NA values
df = df.dropna()

# remove duplicate rows
df = df.drop_duplicates()

# remove where ride_length is 0 or negative
df = df[df['ride_length_min'] > 0.15]
```

Basic Descriptive Statistics & Dealing with Outliers

The analysis of the ride length in the Cyclistic bike-sharing dataset revealed a fascinating insight. The data revealed that the longest recorded ride lasted for an astonishing 34,354.07 minutes, which is nearly 24 days! However, these extreme values are outliers. In general, bike-sharing services such as Cyclistic are meant for short-term transportation, and it is unlikely that customers would use bikes for such a long duration.

These long rides could be indicative of lost or stolen bikes or improperly docked bicycles. Therefore, these extreme values were removed from the dataset by calculating the z-scores. This means that any ride duration that is more than 3 standard deviations (41.32) from the mean has been considered an outlier and removed from the dataset.

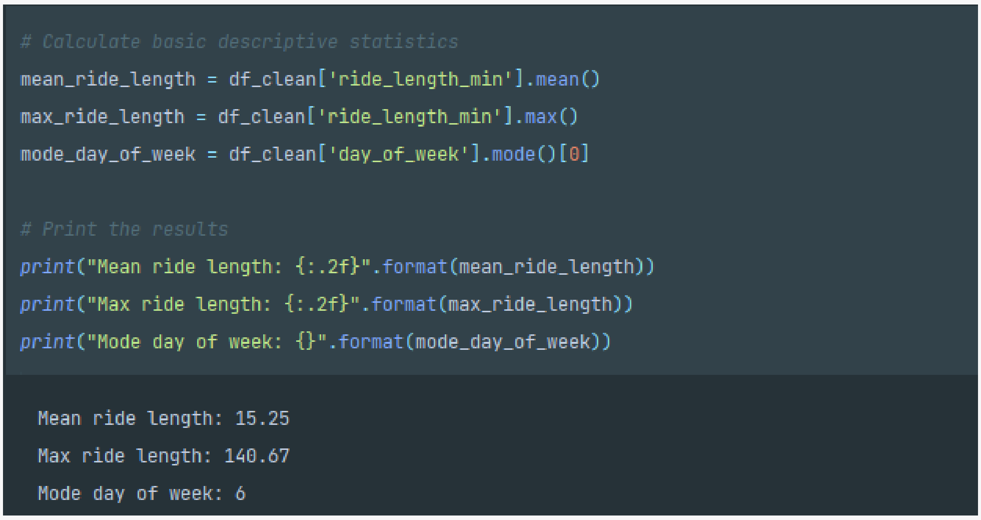


```
# Calculate mean and standard deviation of ride_length_min column
mean_ride_length = df['ride_length_min'].mean()
std_ride_length = df['ride_length_min'].std()

# Calculate z-scores for ride_length_min column
z_scores = np.abs((df['ride_length_min'] - mean_ride_length) / std_ride_length)

# Create new DataFrame without outliers
df_clean = df[z_scores < 3]
```

```
std_deviation = df['ride_length_min'].std()
print("Standard deviation of ride_length_min column: {:.2f}".format(std_deviation))
```

```
Standard deviation of ride_length_min column: 41.32
```

After removing the outliers, the descriptive statistics were recalculated. The mean ride length is now 15.25 minutes, which is slightly lower than the original mean of 16.66 minutes. The maximum ride length is 140.67 minutes, which is much lower than the original maximum of 34,354.07 minutes. The mode day of the week remains the same at 6, indicating that Saturdays are the most popular day for bike rides.

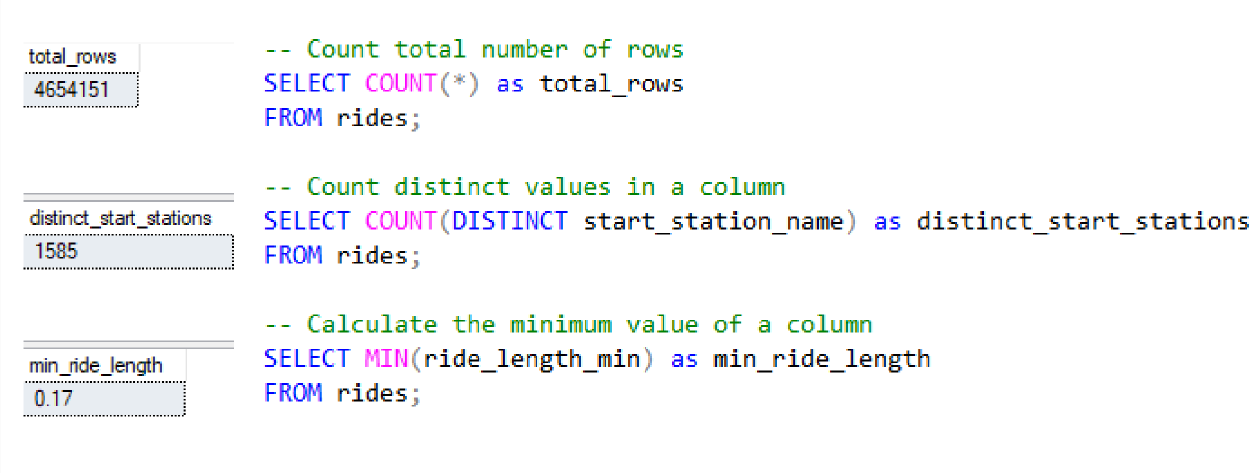


```
# Calculate basic descriptive statistics
mean_ride_length = df_clean['ride_length_min'].mean()
max_ride_length = df_clean['ride_length_min'].max()
mode_day_of_week = df_clean['day_of_week'].mode()[0]

# Print the results
print("Mean ride length: {:.2f}".format(mean_ride_length))
print("Max ride length: {:.2f}".format(max_ride_length))
print("Mode day of week: {}".format(mode_day_of_week))
```

```
Mean ride length: 15.25
Max ride length: 140.67
Mode day of week: 6
```

SQL | Exploring data

Total rows, Distinct Start Stations, Shortest ride



```
total_rows
4654151
```

```sql
-- Count total number of rows
SELECT COUNT(*) as total_rows
FROM rides;
```

```
distinct_start_stations
1585
```

```sql
-- Count distinct values in a column
SELECT COUNT(DISTINCT start_station_name) as distinct_start_stations
FROM rides;
```

```
min_ride_length
0.17
```

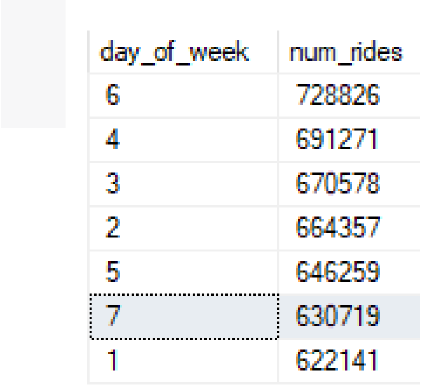
```sql
-- Calculate the minimum value of a column
SELECT MIN(ride_length_min) as min_ride_length
FROM rides;
```

Based on the result of the first SQL query, I can conclude that the dataset contains a total of **4,654,151 rows**, which represents the number of individual rides taken during the specified time period.

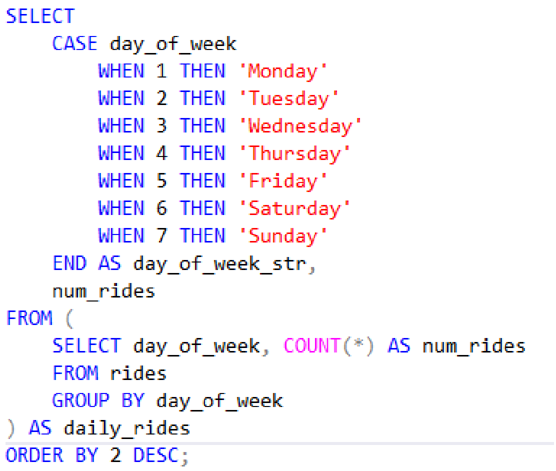
Based on my analysis, I discovered that the Cyclistic service has a total of **1,585 start stations** available for their clients, which provides valuable insight into the breadth and accessibility of the service.

Additionally, I was surprised to find that the minimum ride duration recorded in the dataset was only **17 seconds**, which suggests that some users may be using the service for very short trips or test rides.

Most Popular Weekday



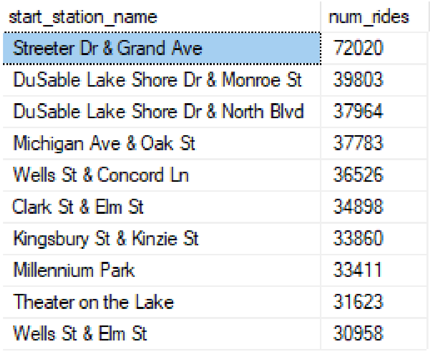
day_of_week	num_rides
6	728826
4	691271
3	670578
2	664357
5	646259
7	630719
1	622141



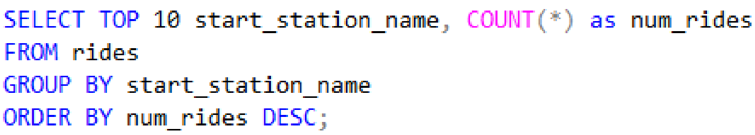
```
SELECT
    CASE day_of_week
        WHEN 1 THEN 'Monday'
        WHEN 2 THEN 'Tuesday'
        WHEN 3 THEN 'Wednesday'
        WHEN 4 THEN 'Thursday'
        WHEN 5 THEN 'Friday'
        WHEN 6 THEN 'Saturday'
        WHEN 7 THEN 'Sunday'
    END AS day_of_week_str,
    num_rides
FROM (
    SELECT day_of_week, COUNT(*) AS num_rides
    FROM rides
    GROUP BY day_of_week
) AS daily_rides
ORDER BY 2 DESC;
```

Based on this query, we can conclude that the **weekends** (Saturday and Sunday) have the **highest number of rides**, with Saturday having the highest number of rides at **728,826**. This suggests that Cyclistic bikes may be popular for leisure and recreational activities on weekends. On the other hand, **Monday has the lowest number of rides at 622,141**, indicating that people may be less likely to use Cyclistic bikes for commuting on Mondays.

Top 10 most popular start stations



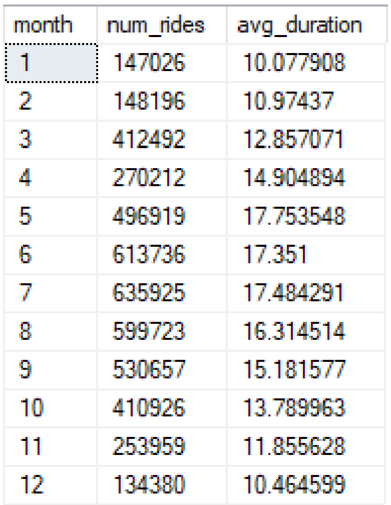
start_station_name	num_rides
Streeter Dr & Grand Ave	72020
DuSable Lake Shore Dr & Monroe St	39803
DuSable Lake Shore Dr & North Blvd	37964
Michigan Ave & Oak St	37783
Wells St & Concord Ln	36526
Clark St & Elm St	34898
Kingsbury St & Kinzie St	33860
Millennium Park	33411
Theater on the Lake	31623
Wells St & Elm St	30958



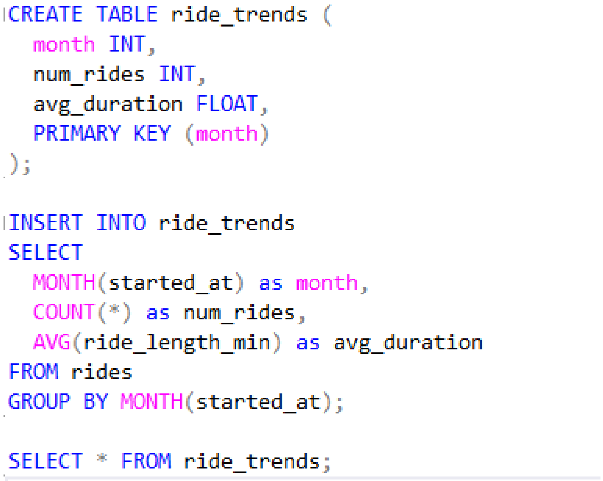
```
SELECT TOP 10 start_station_name, COUNT(*) as num_rides
FROM rides
GROUP BY start_station_name
ORDER BY num_rides DESC;
```

To find out what the top 10 most popular start stations are, I used the SQL query above. The query results show that the station with the highest number of rides is **Streeter Dr & Grand Ave with 72,020 rides**, followed by **DuSable Lake Shore Dr & Monroe St** and **DuSable Lake Shore Dr & North Blvd.**

Trends



month	num_rides	avg_duration
1	147026	10.077908
2	148196	10.97437
3	412492	12.857071
4	270212	14.904894
5	496919	17.753548
6	613736	17.351
7	635925	17.484291
8	599723	16.314514
9	530657	15.181577
10	410926	13.789963
11	253959	11.855628
12	134380	10.464599



```
CREATE TABLE ride_trends (
    month INT,
    num_rides INT,
    avg_duration FLOAT,
    PRIMARY KEY (month)
);

INSERT INTO ride_trends
SELECT
    MONTH(started_at) as month,
    COUNT(*) as num_rides,
    AVG(ride_length_min) as avg_duration
FROM rides
GROUP BY MONTH(started_at);

SELECT * FROM ride_trends;
```

Based on the output, we can see that the number of rides is generally **higher in the warmer months**, with the peak occurring in July. Additionally, the average ride duration tends to increase in the warmest months, with the **highest average duration being in May and June**. This could be indicative of people taking long leisurely rides in the warmer months, while in the colder months, they may be more focused on getting to their destination quickly. There is a **decline in the number of rides from October to November**, which could be due to colder weather and fewer opportunities for outdoor activities.

5. Share

Power BI

In order to effectively communicate my analysis, I used the advanced capabilities of Power BI to craft a comprehensive and user-friendly dashboard. The dashboard is easily accessible via this link.

To create some of the visualisations, I created new columns and measures, which enabled me to perform a more comprehensive analysis of the data. By utilising these techniques, I was able to extract valuable insights from the available data.

5. Act

Final Analysis

1. Annual members represent 61% of the total rides with an average ride duration of 11.88 minutes. Casual members represent 39% with an average ride duration of 20.51 minutes.
2. The most popular bike among the riders was the classic one, however, casual riders have a preference for electric bikes while annual members prefer the classic type.
3. Annual members tend to use the bikes more on weekdays while casual members use them more on weekends. This implies that members use bikes to commute to work on a daily basis while casual riders use them for recreation.
4. We can notice an increase in the number of rides in the warmest months from May to September, this means that the busiest season is the summer, and on the other side, the winter is the season with the lowest number of rides. The main cause for this could be the weather since the cold and rain makes it hard to ride a bike.
5. The busiest start station is at Streeter Dr & Grand Ave, especially for casual riders.

Business Suggestions

After analyse, my suggestions for the marketing team are:

1. As casual riders represent 39% of the total rides and have a higher average ride duration, they could be a valuable target market. Consider **offering promotions** that meet their **preference for electric bikes**, and focus on **promoting weekend** use for recreational activities.
2. Annual members are using the classic bikes more often. We could **promote the benefits of being an annual member**, such as cheaper prices, and advertise the convenience of using bikes for daily commuting during the weekdays.
3. The busiest starting station is at Streeter Dr & Grand Ave, and it's especially popular for casual riders. We could **advertise this station more** and tell people about the fun places they can go to from there.

4. The data indicates that there is an increase in the number of rides during the warmest months of the year, especially from May to September. Consider **offering seasonal promotions** and **incentives to encourage ridership** during the summer months.
5. The data also shows that there is a decrease in the number of rides during the winter months. Consider **offering winter-specific promotions** and **highlighting** the convenience of using bikes for short trips even during colder weather.

Conclusion

This project has been a significant learning experience for me. It has enhanced my analytical abilities and provided me with a deeper understanding of data organization and manipulation techniques using SQL and Python. In addition, I have gained valuable experience working with Power BI, which has allowed me to create insightful visualizations to showcase my findings.

Through this project, I was also able to gain valuable experience in exploring and interpreting data to derive key insights that could answer crucial business questions. This experience has allowed me to become more confident in my ability to work with data and draw meaningful conclusions.